オントロジー対応文法理論と その高速処理のためのコンパイレーション

橋田 浩一・TAM Wailok (産業技術総合研究所)・橋本 泰一 (東京工業大学)・
鷹合 基行 (富士ゼロックス株式会社)・荒牧 英治 (東京大学)

1. はじめに

テキスト入力支援に用いるため、オントロジーと文法に基づいてテキストを修正・補完する技術を開発している[1]。本稿では、そこで用いている形式文法の理論に関して述べ、その文法を文脈自由文法にコンパイルすることによって統語・意味解析および照応解消と談話構造の同定を含む談話解析を高速に実行する方法を論ずる。

2. オントロジー対応文法理論

ISO/TC37/SC4 で橋田をリーダとして策定しつつある。談話構造のアノテーションの方法に関する国際標準 SemAF-DS は、オントロジーのインスタンスである RDF グラフ形式の意味表現を構造記述の一部として生成するような文法に準拠する。この文法は HPSG や LFG で考案された方法をまとめたもので、GDA タグ集合[2]でも用いられている。

この文法は、各文だけでなく、複数文からなる談話にも構造記述を与える。この構造記述は、統語的にはセグメントを節点とする木である。各セグメント *S* の意味構造 *E* は RDF グラフであり、*S* の指示対象を表わす**自己ノード**(self node; 以後 self と略記する)と、原則として *S* の統率子(係り先)の指示対象を表わす**統率子ノード**(governor node; gov と略記)を含む。また、*S* の中の空所の指示対象を表わす**スラッシュノード**(slash node; slash と略記)、*S* より前のセグメントと *S* か *S* の中のセグメントとに共通の指示対象を表わす**後向き中心ノード**(backward-looking center node; bw)と *S* か *S* の中のセグメントと *S* より後のセグメントに共通の指示対象を表わす**前向き中心ノード**(forward-looking center node; fw)を含むことがある。

「身長が 180cm の太郎」の構造記述を図 1 に示す。依存構造を表わす局所木においては、母親の self と gov はそれぞれ主辞の self と gov に等しい。また、原則として、非主辞の娘 *S* の gov と主辞(*S* の統率子)の self とが等しいが、*S* が関係節(たとえば図 1 の Adn)の場合などは、*S* の slash と主辞の自己ノードとが等しい(これは HPSG のスラッシュ導入規則に相当する)。*S* の主辞が他の意味的関係を持つ場合もあるが、本稿では割合する。

等位構造では、等位接続詞の種類により意味構造の作りが異なる。また、等位接続詞以外の娘は slash を共有する。詳細は省略する。

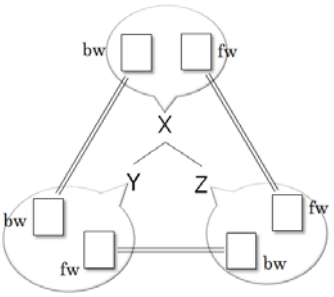

図 2: 局所木における bw と fw の振舞

図 2 に示すように、各局所木においては、母親の bw が左端の娘の bw に伝わり、各娘の fw がその右隣の娘の

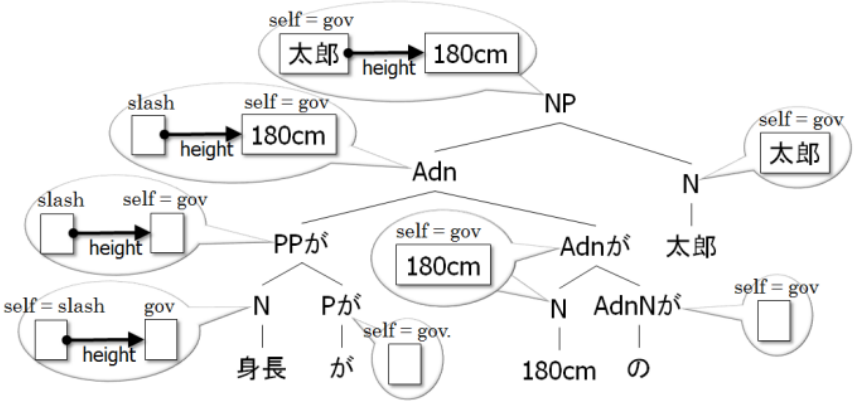

図 1: 「身長が 180cm の太郎」の構造記述

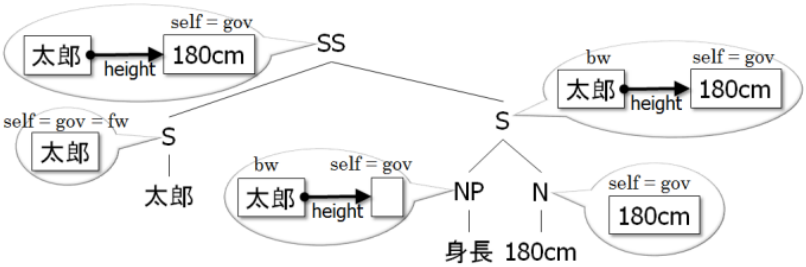

図 3: 文間の照応を含む構造記述

bw に伝わり、右端の娘の fw が母親の fw に伝わる。それによる「太郎。身長 180cm。」という談話の構造記述を図 3 に示す。これは、「太郎」が fw を生成し、それが「身長 180cm」および「身長」の bw に伝わり、それが「身長」に含まれるゼロ照応詞と結び付く様子を表現している。ここでは太郎は「身長」の fw として後続文脈に伝わっていないが、「太郎。身長 180cm。体重 70kg。」のような談話においては伝える必要がある。

この文法では、構造記述の妥当性の一部をオントロジーに基づく妥当性に帰着させる。たとえば図 1 と図 3 では、「身長」や「180cm」の統語範疇を細かく分類するのではなく、height 属性の値である「180cm」というノードのタイプが長さのクラスであるとの意味的制約を用いる。一方、「太郎の身長が 180cm の太郎」などを意味的制約によって排除することはできない。「身長」の必須格がちょうど 1 回だけ埋められるという統語的制約が必要である。具体的には、必須格の補語を要求する「身長」の統語範疇として N 以外のものを設けることになろう。しかし実用的には、そのように統語範疇と文法規則を増やすよりも、統語的制約を単純にして処理の効率を高めた方が良い場合も多いだろう。

3. コンパイレーション

上記のような文法を文脈自由文法にコンパイル(語彙化)することによって高速に処理することが可能と考えられる。そのコンパイレーションは、各セグメントの self と gov のタイプになりうるクラスによってそのセグメントの非終端記号を下位分類し、それに応じて文脈自由規則も下位分類するということである。たとえば「身長」というセグメントの非終端記号は N[Human, Length]のようなものになり、「が」というセグメントの非終端記号は P が[Human, Human]や P が[Length, Length]のようなものになり、それらを含む下記のような文脈自由規則がコンパイレーションにより作られる。

PP が[Length, Length]

→ N[Human, Length]　P が[Length, Length]

Human の下位クラスとしてたとえば Woman が定義されていれば、N[Woman, Length]などの非終端記号とそれらを含む文脈自由規則が作られることになる。

このコンパイレーションには、辞書項目から始めてボトムアップに文法規則をコンパイルして行けば良い。それが収束した後に、談話全体の非終端記号からトップダウンに辿れない無駄な文法規則と辞書項目を捨てる。意味構造は、コンパイル後の文法による解析では作られないが、構文木から明示的に作ることができる。

各セグメントには、slash、bw、fw が各々複数あり得るが、各々の個数に上限を設けることにより、それらを self および gov と同様にコンパイレーションに含めることができる。それでも、slash、bw、fw をコンパイルすべきかどうかは場合によるだろう。これらを含めてコンパイルすることによって文法があまりにも巨大になってしまう場合は、self と gov のみをコンパイルした文法の処理と slah、bw、fw の処理とをリアルタイムに組み合わせた方が良いと考えられる。病理診断報告書のように限定された領域の談話の場合には、slash と bw と fw の値のタイプを限定することができ、これらをコンパイレーションに含めることが現実的である可能性もあるが、その検証は今後の課題である。

4. おわりに

本稿で述べた文法理論に基づいて病理診断報告書のテキスト入力を支援するための文法とオントロジーおよび解析・予測システムを開発している。上記のコンパイレーションを用いた入力支援の実装と評価については機会を改めて報告したい。

謝辞

本研究の一部は、科学技術振興機構 A-STEP ハイリスク挑戦タイプ「自然言語処理とオントロジーに基づく自由テキスト入力支援の医療文書への応用」の一環として実施している。

参考文献

[1] 橋本 泰一・TAM Wailok・鷹合 基行・荒牧 英治・宇於崎 宏・橋田 浩一 (2011) 病理診断報告書作成のためのオントロジーを利用したテキスト入力支援. 言語処理学会第 17 回年次大会, 940-943.

[2] 橋田 浩一 (2005) GDA 日本語アノテーションマニュアル. http://i-content.org/gda/tagman.html